Exercise 1: Entropy

A fair die is rolled at the same time as a fair coin is tossed. Let A be the number on the upper surface of the dice and let B describe the outcome of the coin toss, where

$$B = \begin{cases} 1, & \text{head}\,, \\ 0, & \text{tail}\,. \end{cases}$$

Two random variables X and Y are given by $X = A + B$ and $Y = A - B$, respectively.

(a) Calculate the entropies $H(X)$ and $H(Y)$, the conditional entropies $H(Y|X)$ and $H(X|Y)$, the joint entropy $H(X, Y)$ and the mutual information $I(X; Y)$.

(b) Show that, for independent discrete random variables X and Y,

$$I(X; X + Y) - I(Y; X + Y) = H(X) - H(Y)$$

Exercise 2: Mutual Information of Three Variables

Let X, Y, and Z be three discrete random variables. The mutual information of X, Y, and Z is defined as:

$$I(X;Y;Z) = \sum_x \sum_y \sum_z p(x,y,z) \log\left(\frac{p(x,y)p(x,z)p(y,z)}{p(x)p(y)p(z)p(x,y,z)}\right). \tag{1}$$

(a) Prove the lemma: $I(X;Y;Z) = I(X;Y) - I(X;Y|Z)$. Note that the conditional mutual information is defined as:

$$I(X;Y|Z) = \sum_z \sum_x \sum_y p(z)p(x,y|z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)}. \tag{2}$$

(b) Prove the following relation with the above lemma:

$$I(X;Y) = I(X;Y|Z) + I(Y;Z) - I(Y;Z|X). \tag{3}$$

Exercise 3: Smoothed Cross-Entropy Loss

Over-confidence is a state when a model is more confident in its prediction than the input data warrants. Label smoothing (a.k.a. smoothed cross-entropy loss) [1] is a widely used trick in deep learning classification tasks for alleviating the over-confidence issue and increasing model robustness. In the conventional cross-entropy loss, we aim to minimize the KL-divergence between d and $\pi(\mathbf{x}|\theta)$, where the ground truth distribution d is a delta-distribution (i.e., only $d_k = 1$ for the ground truth class), and $\pi(\mathbf{x}|\theta)$ is the predicted distribution by the model π parameterized by θ. The key step in label smoothing is to smooth the ground truth distribution. Specifically, given a hyper-parameter β (e.g., $\beta = 0.1$), we uniformly distribute the probability mass of β to all the g classes and reduce the probability mass of the ground truth class. Consequently, the smoothed ground truth distribution $\tilde{d}$ is

$$\tilde{d}_k = \begin{cases} \frac{\beta}{g} & \text{for } d_k = 0; \\ 1 - \beta + \frac{\beta}{g} & \text{for } d_k = 1. \end{cases} \tag{4}$$

The smoothed cross-entropy is then $D_{KL}(\tilde{d}||\pi(\mathbf{x}|\theta))$.

(a) Derive the empirical risk when using the smoothed cross-entropy as loss function. (Hint: some terms can be merged into a constant and ignored during implementation).

(b) Implement the smoothed cross-entropy. We provide the signature of the function here as a reference:

```r
#' @param label ground truth vector of the form (n_samples,).
#'  Labels should be "1","2","3" and so on.
#' @param pred Predicted probabilities of the form (n_samples,n_labels)
#' @param smoothing Hyperparameter for label-smoothing

smoothed_ce_loss <- function(
label,
pred,
smoothing){
  return (loss)
}
```

References

[1] Szegedy, Christian, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. ”Rethinking the inception architecture for computer vision.” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818-2826. 2016.